Improving the Predictive Validity of Reading Comprehension Using Response Times of Correct Item Responses

Shiyang Su

University of Central Florida

Mark L. Davison

University of Minnesota

Su, S., & Davison, M. L. (2019). Improving the Predictive Validity of Reading Comprehension Using Response Times of Correct Item Responses. *Applied Measurement in Education, 32*(2), 166–182. https://doi.org/10.1080/08957347.2019.1577247
Published online: 13 Mar 2019

Correspondence concerning this manuscript should be addressed to Dr. Shiyang Su, Department of Psychology, University of Central Florida, 4111 Pictor Lane, Orlando, Florida, 32816.

EMAIL: Shiyang.Su@ucf.edu

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A140185. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Improving the Predictive Validity of Reading Comprehension Using Response Times of Correct Item Responses

Abstract

Response times have often been used as ancillary information to improve parameter estimation. Under the dual processing theory, assuming reading comprehension requires an automatic process, a fast, correct response is an indicator of effective automatic processing. A skilled, automatic comprehender should be high in response accuracy and low in response times. Following this argument several questions were addressed in this study. First, individuals with higher ability endorsed a correct response more quickly in the reading comprehension assessment, suggesting *correct* response times provide useful information in discriminating individuals of different abilities. Second, in terms of predicting an external criterion of reading proficiency, the increment of predictive validity of categorized *correct* response times was larger than the incremental validity of continuous response times based on both correct and incorrect response times. An index reflecting both accuracy and response time yielded higher incremental validity than indices reflecting response time only. Results support response time as manifesting a dimension of interest in its own right, over its inclusion as an ancillary dimension in a multidimensional model.

Introduction

With online assessment becoming mainstream and the recording of response times (or latencies) becoming straightforward, the importance of response times as a measure of psychological constructs has been recognized and the literature of modeling response times has been growing during the last few decades. Previous studies have tried to formulate models and theories to explain the construct underlying response times, to examine the relationship between response times and response accuracy, and to understand individuals' behaviors and cognitive processes implied by response times. Quite a few studies suggest that response times can be used as an additional source of information to improve the precision in estimating person ability parameters and item parameters (Ferrando & Lorenzo-Seva, 2007; Molenaar, Tuerlinckx, & van der Maas, 2015; Petscher, Mitchell, & Foorman, 2015; Thissen, 1983; van der Linden, 2007; Wise & DeMars, 2006). Studying response times could advance our understanding of the implicit cognitive processes and provide insights into cognitive theories (Egloff & Schmukle, 2002; Partchev & De Boeck, 2012; Petscher et al., 2015). Given the convenience of obtaining time data in computerized assessments, and the potential gain in precision with no additional cost, it becomes intensely popular to develop approaches taking response times into account.

Much of the research to date has been devoted to the development of models. At some point, however, the field must move beyond modeling to consider the measurement properties of scores derived from response time models, reliability and especially validity. Broadly speaking, there are two possible uses for a response time dimension. First, a response time dimension may be conceived as a useful ancillary dimension that is not of interest in its own right, but whose inclusion in multidimensional models improves the reliability of a second dimension in the model (Petscher et al., 2015). For instance, the response time dimension may be an ancillary

dimension whose inclusion in a multidimensional model controls error in the measurement of achievement resulting from individual differences in the speed-accuracy trade-off. Alternatively, the response time dimension may be conceived as a target dimension of interest in its own right (Davison, Semmes, Huang, & Close, 2012; Semmes, Davison, & Close, 2011). In this study, we conceptualized the response time dimension as a target, efficiency dimension that could improve the prediction of future performance over and above an achievement test score alone. The dimensions derived from two different response time models were compared in terms of their contributions to predictive validity in an attempt to evaluate the conceptualization of a response time dimension as useful in its own right.

Besides the issue of response time as an ancillary variable or a target variable, there is the issue of whether to model a single dimension that reflects both speed and accuracy (Dennis & Evans, 1996; Maris & van der Maas, 2012), to model speed and accuracy as separate dimensions, or some combination of the two. If both, then the dimension reflecting both speed and accuracy is a higher order dimension with separate speed and accuracy dimensions at the lower order. In our validity analyses, we included an accuracy (or ability) dimension, two speed dimensions, and a dimension that reflected both speed and accuracy.

Response Times

Different models have been proposed to model response times in the psychometric, behavioral, and cognitive psychology literature. The most widely used approach in the recent literature is to model response times and responses jointly. Thissen (1983) was among the first who modeled the response times and responses jointly with an item response theory (IRT) framework. Assuming the random variable of response times follows a lognormal distribution, the log transformed response time (logtime) was formulated to link both a person slowness (i.e.,

speed) parameter and an item slowness parameter. However, Thissen's (1983) model didn't contain a time discrimination parameter, therefore allowing no difference in time discrimination across items. The hierarchical framework of modeling response times and responses derived by van der Linden (2007) is another frequently cited model. On the first level, it models responses exclusively dependent on latent ability in the IRT model, and response times exclusively dependent on latent speed in the lognormal model; on the second level, the joint, population-wise distributions of IRT and time parameters are considered to represent the relation between speed and ability. Following this argument, many studies derived innovative models based on the hierarchical framework (e.g., Loeys, Rosseel, & Baten, 2011; Wang & Xu, 2015). Specifically, Molenaar, Tuerlinckx and van der Maas (2015) re-formulated van der Linden's (2007) model so that the linear factor model of logtime links both functions of response times and responses. The hierarchical model is rewritten as an oblique two-factor model with dichotomous indicators for the ability factor and continuous indicators for the speed factor that are linked by cross-relations.

As suggested by previous studies, one benefit of using response times is to enhance the psychometric properties of a test. Different estimates of reliability and precision have been used in previous studies. Ferrando and Lorenzo-Seva (2007) modified Thissen's (1983) model and assessed the precision of person parameter estimates (i.e., expected a posteriori) of a personality measure through the information function, marginal reliability and root mean square error (RMSE), and observed a slight increase in precision, more at the extreme ends of the theta continuum, when item response times were taken into consideration. Van der Linden (2006) used the data from a computerized adaptive test of arithmetic reasoning to assess the validity of the lognormal model of response times using model fit statistics and posterior predictive checks on observed and expected response times, and suggested the lognormal model had a superior

performance to the normal alternative. Using data of 3rd graders, Petscher et al. (2015) evaluated the differential reliability of ability estimates in a vocabulary knowledge test which captured both responses and response time using van der Linden's (2007) hierarchical model, and compared ability estimates using classical test theory and IRT. Results indicated that compared with the usual IRT model, the ability estimates benefited from the hierarchical model including both responses and response times, with an average of 0.05 increase in the reliability and reduction in the standard error. The improvements were greater for individuals with high or average ability, whereas low-ability individuals gained little from including response times. Motivated by this finding, the current study is interested in if using response times of *correct* responses would benefit the person ability estimates as much as using response times of all responses. The authors have completed a study assessing the utility of modeling correct response times and observed better model fit statistics, higher marginal reliability and information functions. The focus of the current study is to compare with approaches using response times of all responses and expect the gain is no less when using response times of correct responses.

Even though a few studies have assessed the improvement of criterion-related validity of incorporating response times, not all of these are in the achievement domain. In the anxiety literature, a study by Egloff and Schmukle (2002) used response times as part of their measure of implicit anxiety and observed compelling evidences of predictive validity for this new measure through the prediction of several relevant criteria. In a study of an achievement measure, Wise and DeMars (2006) found that compared with the standard 3PL model, an effort-moderated model reflecting examinee effort indicated by response times fitted the examinees' response patterns better, provided more accurate parameter estimates, and showed higher convergent validity through correlations with external criteria including SAT-Verbal, SAT-Quantitative and

GPA. In the chess domain, Molenaar et al. (2015) evaluated the increments of predictive validity when using response times as ancillary information. Using a "gold standard" for chess ability as the criterion, they compared the performance of latent trait estimates from the two-parameter logistic (2PL) model without response times and generalized linear IRT based models including response time parameters, and observed increments of R^2 and model fits. Nevertheless, the observed increments in precision and validity were usually not substantial in previous studies. For example, the increment of R^2 in predicting an external criterion using the hierarchical model was only 0.01 – 0.02 compared against the model without response times (Molenaar et al., 2015). According to the authors' knowledge no studies have examined concurrent, predictive, or incremental validity of response time in a reading comprehension assessment through validation with an external criterion.

Another issue of interest is the correlation between latent speed and latent ability. Even though recent literature frequently proposes joint modeling of speed and accuracy based on response times and responses, the observed correlation between latent speed and latent ability ranges from close to zero (Goldhammer, Naumann, Stelter, Tóth, Rölke, & Klieme, 2014; Partchev, De Boeck, & Steyer, 2013; van der Linden, Scrams, & Schnipke, 1999), to moderate (Petscher et al., 2015; van der Linden, 2007), or sometimes negative (van der Linden & Guo, 2008). According to van der Linden (2007; 2009), speed and accuracy are conditionally independent within person but can be correlated between persons. The impression of these studies is that the size and sign of correlation between speed and ability might depend largely on the nature of the test and test takers' time-management strategies, although a large body of studies observed a positive-but-weak correlation (van der Linden, 2009). Using a dual processing theory, Goldhammer et al. (2014) further stated that the direction of relation between speed and

ability depends on the cognitive processing required by the test: in tests requiring controlled processing (e.g. problem solving tasks), increments in response times are associated with higher ability, whereas in tests requiring automatic processing, such as reading literacy assessment in their example, increments in response times are associated with lower ability; in other words, working automatically, and thus faster, reflects higher ability in reading tasks. Some or all of the reading comprehension process becomes less conscious and less deliberate, namely, *automatic*, as the comprehension processes develop. Fast speed of endorsing *correct* responses in a reading test is an indication of such automatic processing. It is difficult to suppress or to alter an automatic process once learned (Goldhammer et al., 2014). Petscher et al.'s (2015) study of reading also suggested that higher ability individuals (3^{rd} graders in their example) tend to spend less time per item than lower ability individuals. Therefore, the current study is interested in examining whether the individuals with higher ability endorse a *correct* response more quickly than lower ability individuals, in the reading comprehension assessment. The way response times being utilized provides insights into this implicit cognitive process. If development of reading skills through elementary school substantially increases *automaticity* in processing word units into recognizable words and connecting words when reading a passage, it is particularly interesting to examine the dual process of reading comprehension for elementary students.

Most studies mentioned above used continuous response times for all responses (both correct and incorrect) and assumed a lognormal distribution of response times. There were also a few studies treating response times as categorical variables. For example, when an abnormal behavior such as rapid guessing was present, a dichotomous index that was reflected by response times could be included in the model to indicate examinees' behavior (e.g., effort; Wise & DeMars, 2006). Besides differentiating patterns of behaviors, response times were also used as

dichotomous variables to imply types of cognitive process. Motivated by the dual processing theory, Partchev et al. (2013; De Boeck & Partchev, 2012) treated the response time as a dichotomous variable by categorizing the item response times into fast and slow using the empirical median as the cut-off. They indicated that fast correct responses and slow correct responses involved different processes and abilities in a matrices test and a verbal analogies test. Their studies demonstrated how cognitive theory and measurement model can be integrated to model the sequential process of automatic responses. However, it remains a question whether the categorized response times of correct responses can be as useful as the continuous response times of all responses (both correct and incorrect) in terms of predicting an external criterion.

Model Comparison

The usefulness of categorized, correct response times in estimating the latent trait of a reading comprehension test is assessed through prediction toward an external measure of reading proficiency. Latent trait estimates from the models of categorized correct response times are compared with those from the model of response accuracy and those from the hierarchical model of continuous response times.

Unidimensional Accuracy Model. The measurement model of response accuracy is a unidimensional 2PL model using only response data. The probability of responding correctly (i.e., $X_{j1} = 1$) is posited as

$$\pi\left(X_{j1} = 1 \middle| \theta_a, \alpha_{j1}, b_{j1}\right) = \frac{\exp[\alpha_{j1}(\theta_a - b_{j1})]}{1+\exp[\alpha_{j1}(\theta_a - b_{j1})]}, \quad (1)$$

where α_{j1} is the discrimination parameter for item j, b_{j1} is the difficulty parameter for item j, and θ_a in this model is interpreted as the latent ability of reading comprehension: examinees with high θ_a estimates are described as good comprehenders and those with low θ_a estimates are

described as poor comprehenders. In Equation (1) respondents' ability is an exclusive function of response accuracy.

Unidimensional Efficiency Model. The dichotomized, correct response time variable X_{j2} denotes the speed of choosing a correct response, which we call the comprehension efficiency (or rate) dimension. $X_{j2} = 1$ if a correct response was chosen in a fast way, $X_{j2} = 0$ if a correct response was chosen slowly, and X_{j2} is missing if an incorrect response was chosen. The probability of $X_{j2} = 1$, conditional on $X_{j1} = 1$, takes the 2PL form

$$\pi\left(X_{j2} = 1 \middle| X_{j1} = 1, \theta_e, \alpha_{j2}, b_{j2}\right) = \frac{\exp[\alpha_{j2}(\theta_e - b_{j2})]}{1+\exp[\alpha_{j2}(\theta_e - b_{j2})]}, \quad (2)$$

where α_{j2} is the discrimination parameter specific to an efficient process of solving item j, b_{j2} is the difficulty parameter specific to an efficient process of solving item j. θ_e is the latent trait of comprehension efficiency, which is interpreted as the propensity of choosing a correct response fast over choosing a correct response slowly. A high θ_e estimate indicates a tendency of comprehending efficiently and a low θ_e estimate indicates a tendency of comprehending inefficiently. The response time is conceptualized as a target, efficiency dimension.

Two-Dimensional Tree Model. In the condition that θ_a and θ_e are two distinct but related latent traits, we use a two-dimensional 2PL IRT tree model (2DTREE; De Boeck & Partchev, 2012) and estimate the correlation between the two dimensions. We denote the corresponding latent traits in 2DTREE model as $\theta_{a.2D}$ and $\theta_{e.2D}$. The multidimensional model assumes that each of the two dimensions, manifested in either response accuracy or response times respectively, also provides ancillary information for the second dimension.

Unidimensional Automaticity Model. Based on the dual processing theory, we model automaticity as a single dimension that reflects both speed and accuracy (Goldhammer et al.,

2014). $X_j = 2$ if item j was endorsed correctly and fast, $X_j = 1$ if item j was endorsed correctly but slow, and $X_j = 0$ if item j was endorsed incorrectly. The probability of the unidimensional, ordinal polytomous model of automaticity can be written in terms of conditional probabilities in Equations (1) and (2) as:

$$\pi(X_j = 0) = 1 - \pi(X_{j1} = 1), \quad (3)$$

$$\pi(X_j = 1) = \pi(X_{j1} = 1)[1 - \pi(X_{j2} = 1|X_{j1} = 1)],$$

$$\pi(X_j = 2) = \pi(X_{j1} = 1)[\pi(X_{j2} = 1|X_{j1} = 1)].$$

Compared with other polytomous models, the Graded Response Model (GRM, Samejima, 1969) was preferred due to its advantages in reliability, model fit, and information function when measuring comprehension automaticity (Su, 2017). Therefore, the GRM was implemented to assess comprehension automaticity in the present study. The probability of responding with a given category is obtained by subtraction of adjacent boundary functions,

$$\pi(X_j = c|\theta_{au}, \alpha_j, \delta_{jc}) = \frac{\exp[\alpha_j(\theta_{au} - \delta_{jc})]}{1+\exp[\alpha_j(\theta_{au} - \delta_{jc}]} - \frac{\exp[\alpha_j(\theta_{au} - \delta_{j(c+1))}]}{1+\exp[\alpha_j(\theta_{au} - \delta_{j(c+1))}]}, \quad (4)$$

where θ_{au} is the latent trait of comprehension automaticity, δ_{jc} and $\delta_{j(c+1)}$ are the location parameters for category c and category c+1 of item j respectively.

Hierarchical Model. As a standard to compare with the performance of the proposed approaches of using categorized correct response times in our validity analyses, van der Linden's (2007) hierarchical model of continuous response times was implemented. Two measurement models are specified at the first level. The measurement model of response accuracy is a function of latent ability, represented in Equation (1). The measurement model of response times is a function of latent speed. The logarithmic transformation was applied to the raw response times of all responses, since the lognormal distribution has been widely assumed when modeling

continuous response times (Ferrando & Lorenzo-Seva, 2007; Thissen, 1983; van der Linden, 2007). Fox, Klein Entink, & van der Linden (2007) used the specification of a lognormal model to link the observed log response times to latent speed variable as below:

$$lnt_{ij} = \lambda_j + \varphi_j\tau_i + \omega_{ij} \quad (5)$$

where φ_j is a time discrimination parameter for item j, λ_j is a time intensity parameter for item j, τ_i is the latent speed variable for person i. Due to the positive sign before τ_i, this speed variable is indeed an indicator of *slowness*: a higher value of latent speed variable τ_i denotes a lower speed in practice. The random error term ω_{ij} of the logtime $\ln t_{ij}$ is assumed to follow a normal distribution, which implies that the model belongs to a lognormal family. At the second level, the two measurement models are connected by correlating item and person parameters.

Molenaar et al. (2015) re-formulated the hierarchical model (Fox et al., 2007; van der Linden, 2007) to a generalized linear latent variable model for the analysis of responses and response times, so that it could be directly implemented in Mplus. The hierarchical model could be written as separate generalized linear measurement models of responses and response times that are linked by cross-relations. The re-formulated model leaves the measurement model of response accuracy untouched, while adding a component in the measurement model of response times. Accordingly, the measurement model of response accuracy is a function of $\theta'_{a,i}$, identified by Equation (1), whereas the measurement model of response times is specified as:

$$lnt_{ij} = \lambda_j + \varphi_j\tau'_i - \varphi_j\rho\theta'_{a,i} + \omega_{ij}. \quad (6)$$

The latent traits, τ'_i and $\theta'_{a,i}$, are assumed uncorrelated at first; then the latent correlation, denoted by ρ, between the latent speed (or slowness) τ'_i and the latent ability $\theta'_{a,i}$ is estimated. A comparison of Equations (5) and (6) allows us to rewrite $\tau_i = \tau'_i - \varphi_j\rho\theta'_{a,i}$. When $\rho = 0$, $\tau_i = \tau'_i$. Equation (6) leaves the measurement model for responses intact, and models the

information about $\theta'_{a,i}$ that is available in the response times (if any). This requires a cross-relation function in the measurement model of response times, but not in the measurement model of response accuracy. To put Equation (6) on the same scale of the original hierarchical model of van der Linden (2007), a constraint of latent trait variance, $\sigma^2_{\tau'} = 1 - \rho^2$, is needed to identify τ'.

Research Questions

In summary, the literature to date raises several important questions that go well beyond modeling. First, should we conceptualize the response time dimension as a useful ancillary variable or as a target variable of interest in its own right? Secondly, should item response times be conceptualized as continuous variables defined over all responses, both correct and incorrect, or should they be conceptualized as dichotomous variables (fast vs. slow) and defined only over correct responses? Finally, should our theories include an efficiency (or speed) construct manifested only in response times, an automaticity construct manifested in both response accuracy and response times, both, or neither? No one study can fully address all of these issues, but the current study presents some evidences based on real data relevant to each of these issues.

The following study began by deriving estimates of latent ability with and without including response time as an ancillary variable to assess whether including the ancillary variable yields a superior estimate of ability that discriminates the criterion variable. Second, measures of efficiency based on response time alone were computed to assess whether models with and without efficiency predicted the criterion variable equally well. Here, we compared two regression models, one in which the dimension was based on continuous response times of both correct and incorrect responses, labeled speed, and one in which the dimension was based on dichotomized, correct response times, labeled efficiency. Finally, we fit models with and without

an automaticity dimension to see if inclusion of the automaticity dimension improved prediction. We also compared the incremental validity of each new variable to that of the ability variable.

Method

Sample

A sample of 4,432 students in Grades 3-5 from over 50 schools in 13 states participated in the *Multiple-Choice Online Causal Comprehension Assessment* (MOCCA; Carlson, Seipel, & McMaster, 2014) study in 2016. The sample was recruited through online solicitation of schools. Three different forms were administrated to each grade. After cleaning, data of 4,288 students were used in the study. Table 1 shows the sample sizes of each form and aggregated by grade.

Following MOCCA testing, students also take the statewide assessment: Smarter Balanced Assessment Consortium (SBAC; 2016), which is an untimed test in Mathematics and English Language Arts/Literacy. The SBAC Technical Manual (2016) reports estimated marginal reliabilities for the overall English Language Arts and Mathematics tests of .91 or higher for all Grades 3-5. Students are classified as proficient or non-proficient in the SBAC test, where proficient students demonstrate the knowledge and skills necessary for college and career readiness. To evaluate the predictive validity of the MOCCA, we collected 1,057 students' SBAC proficiency classifications in reading (i.e., English Language Arts/Literacy). Table 1 shows the sample sizes for those having SBAC classifications at each grade. SBAC proficiency classifications, rather than scale scores, were available for a larger sample of students.

Reading Comprehension Test

MOCCA is a 40-item multiple-choice, causal comprehension assessment which was designed to identify comprehension processes used during reading of narrative texts in Grades 3-5 (Carlson et al., 2014). There are nine forms. Students took the MOCCA online test with

computers or tablets in classrooms or in the school computer labs, under the supervision of a teacher or other school staff. Each form has one version of forward order of items and another version of backward order of items. Participants are randomly assigned to one of six versions (3 forms x 2 item orders) of the test at their grade. An item is a short story, containing a title and seven sentences with the 6th sentence deleted. Participants are required to choose one of three alternative response types to fill in the deleted sentence. There is only one question per story passage. The correct answer represents a causally coherent inference, which closes the causal gap between the 5th and 7th sentences. The test itself imposes no time limit, but testing typically occurred during one class period (i.e., 30-60 min) , with a mean student testing time of 35 minutes across grades. Most students can complete the test without rushing. Students with less than ten responses (1.67%) were eliminated from the sample.

Analysis

In order to compare estimates from different models, all models were implemented in Mplus (Muthen & Muthen, 1998-2015), with the maximum likelihood estimation (MLE) using the Expectation-Maximization (EM) algorithm (Bock & Aitkin, 1981). With the full sample, for each person we obtained estimates of latent traits, including θ_a in Equation (1), θ_e in Equation (2), $\theta_{a.2D}$ and $\theta_{e.2D}$ in the 2DTREE model, θ_{au} in Equation (4), τ in Equation (5) (without considering the hierarchical relation with latent ability on the second level), and θ'_a and τ' in Equation (6). Equation (1) is the response-accuracy-only model. Equations (2) and (5) are the response-time-only models. Equations (3) and (6) are the models considering both response accuracy and response times. We assessed the correlation between estimates of these latent traits. In the result section, we use the term "ability" to indicate a latent trait manifested in response accuracy, including θ_a, $\theta_{a.2D}$ and θ'_a; the term "efficiency" to indicate a latent trait manifested

in categorized response times of correct responses, including θ_e and $\theta_{e.2D}$; the term "speed" to indicate a latent trait manifested in continuous response times of all responses, including τ and τ'; the term "automaticity" to indicate θ_{au}, a latent trait manifested in both response accuracy and response times. Note that the latent speed was scored in an opposite direction from the latent efficiency that a fast respondent should receive a low score of latent speed.

To evaluate the predictive validity of the measures, students' SBAC reading proficiency classification in the reduced sample was used as the external criterion to assess its association with estimates of automaticity and efficiency above and beyond ability estimates of response accuracy. As a comparison, the latent speed in the well-established hierarchical model was used as an additional predictor to evaluate its contribution above and beyond the latent ability in predicting the criterion. Increments in R^2, models fit statistics and deviance tests are reported as validity evidences. Also, the areas under the ROC curves (AUC) are calculated and reported to distinguish proficient readers and non-proficient readers. Higher AUC scores indicate better prediction, and a score ≥ 0.8 is considered good discrimination.

Results

Response Times for Correct and Incorrect Responses

Figure 1 shows the average log response times (i.e., logtimes) per item for correct responses and incorrect responses for each form of Grade 4 as an example. For each grade, the average logtimes for correct responses are consistently and significantly higher ($p < .01$), and less variant than the average logtimes for incorrect responses. Given the difference in correct and incorrect responses, an ANOVA model was used to test the effect of response accuracy on logtimes after controlling for item and person. With logtimes being the dependent variable, the ANOVA model included accuracy (i.e., correct or incorrect) as crossed with item and with

person. For each form, the *F* test suggests a significant ($p < .01$) main effect of response accuracy on logtimes after controlling for item and person. The results suggest that if speed is measured by response times of both correct and incorrect responses, then low achieving students would have faster average speed.

Response Times for High and Moderate Ability Students

We grouped respondents by the latent ability (θ_a) estimated as an exclusive function of response accuracy in Equation (1), such that the high θ_a respondents are defined as those whose θ_a's are above 75th percentile and the moderate/medium θ_a respondents are defined as those whose θ_a's are within 50th-75th percentile. For each form in Grades 3-5, respondents with high θ_a tend to spend significantly less time ($p < .01$) to answer an item correctly for most of the items than those with moderate θ_a. This is consistent with Petscher et al.'s (2015) finding that low and average ability individuals had an approximately equal frequency of being fast or slow, whereas the high ability individuals maintained a stronger representation of fast speed.

To further examine the difference of average log response times of *correct* responses (i.e., correct logtimes) between respondents with high and moderate θ_a's, an ANOVA model was used to test the effect of θ_a level on correct logtimes after controlling for item and person. With correct logtimes being the dependent variable, the ANOVA model included ability levels (i.e., high or moderate θ_a) and item as crossed, and person nested within ability levels. For each form, the *F* test suggests a significant main effect ($p < .01$) of ability levels on correct logtimes after controlling for item and person.

As a comparison, we also assessed the logtimes of incorrect responses per item for high θ_a respondents and moderate θ_a respondents for each form in Grade 3-5. With incorrect logtimes being the dependent variable, the ANOVA model included ability levels and item as crossed, and

person nested within ability levels. There is no significant difference between incorrect logtimes of the two ability levels after controlling for item and person for all forms except for Form 4.2.

Relations of Ability, Speed, Efficiency, and Automaticity

Table 2 shows the correlation between latent trait estimates in different models. The latent abilities in the accuracy only model (θ_a in Equation 1), the hierarchical model (θ'_a in Equation 6), and 2DTREE model ($\theta_{a.2D}$) are perfectly correlated with each other. This raises doubt as to whether including speed as an ancillary variable improves or even changes the estimate of ability. The correlation between θ_e and its counterpart $\theta_{e.2D}$ is also perfect. τ is strongly correlated with its counterpart τ'. The correlation between latent speed and latent ability is moderate to low. This suggests that the speed and efficiency variables provide information that differs from that in the ability variables and supports the discriminant validity of those variables. The correlation between latent speed and latent efficiency is moderate in the negative direction, but only because the two dimensions are reverse scored. The absolute values of the correlations are not so high as to suggest that speed, based on response times measured continuously for all responses (both correct and incorrect), is equivalent to the efficiency variable based on *correct* response times scored dichotomously as fast vs. slow.

Automaticity θ_{au} is more highly correlated with ability θ_a than with speed τ or efficiency θ_e. Even though it is more highly correlated with ability, they are not so highly correlated that one could consider ability and automaticity to be the same variable. The correlations range from .80 to .86 across forms and grades.

Figures 2-4 show the scatter plots of estimates of different latent traits for Grade 4 as an example, since the patterns are similar across grades. Figure 2 compares the ability estimates θ_a, θ'_a and $\theta_{a.2D}$ with the automaticity estimates θ_{au} of the GRM. The scatter plots provide

information about where the disparity of ability estimates and automaticity estimates is. The estimates are scattered on the high end of the theta scale. On the low end of the theta scale, the ability estimates are highly overlapped with the automaticity estimates. Since both the correlation and scatter plots show that the estimates of θ'_a of the hierarchical model and $\theta_{a.2D}$ of the 2DTREE model are perfectly overlapped with θ_a of the unidimensional model, in Figure 3 only the ability estimates θ_a are compared with the estimates of speed and efficiency respectively. There is almost no association between θ_a and θ_e or $\theta_{e.2D}$. The relation between estimates of ability θ_a and speed τ depends on θ_a: for θ_a above 0, τ stays consistent across θ_a; for θ_a below 0, τ varies extensively regardless of θ_a. When θ_a is below 0, similar to the relation between θ_a and τ, τ' varies extensively regardless of θ_a; when θ_a is above 0, τ' decreases gradually as θ_a increases, suggesting a faster speed with higher ability. It is consistent with the ANOVA result that high ability respondents require less time to answer an item correctly than moderate ability respondents. Figure 3 also shows that the variability of speed in low ability students is much higher than the variability of speed in high ability students. This pattern is consistent with Figure 1 that the speed of endorsed correct responses is less variant than the speed of endorsed incorrect responses. Figure 4 compares efficiency and speed. We observe that speed estimates are highly correlated with efficiency estimates; while in the upper interval of efficiency dimension, the discrepancy between speed and efficiency starts to emerge.

Table 3 reports the scaled conditional standard error of ability dimensions averaged over examinees across the theta interval. To scale the conditional standard error, the standard error of ability estimate of each person is divided by the standard deviation of the ability dimension. The average scaled conditional standard errors of different ability dimensions are not significantly different from each other. These results suggest that including speed or efficiency as an ancillary

variable did not have much effect on the standard errors of ability. The results in Tables 2 and 3 raise doubts about the utility of speed or efficiency as ancillary variables, because the measurements of accuracy were virtually the same with similar standard errors with or without including speed or efficiency in the model.

Criterion-related Validity of Accuracy, Speed, Efficiency, and Automaticity

To evaluate the criterion-related validity of the measurements, students' SBAC reading proficiency classification in the reduced sample was used as the external criterion to assess its association with estimates of efficiency, speed, and automaticity above and beyond ability estimates manifested in response accuracy. It is important to note that the SBAC proficiency classification is a binary variable (1= proficient, 0=non-proficient) and is based purely on response accuracy of an untimed test. 49% students in Grade 3 are classified as proficient and 51% as non-proficient; 52% students in Grade 4 are classified as proficient and 48% as non-proficient; 53% students in Grade 5 are classified as proficient and 47% as non-proficient. Table 4 shows the average estimates of non-proficient and proficient students respectively in Grades 3-5. Estimates of ability dimensions and automaticity are significantly different between proficient and non-proficient readers across grade, suggesting the new test is a valid measure of reading comprehension. Estimates of efficiency and speed are significantly different between proficient and non-proficient readers in Grades 4 and 5, but not in Grade 3. Proficient readers in Grades 4-5 also display significantly faster speeds (lower scores of τ/τ', higher scores of $\theta_e/\theta_{e.2D}$). It is likely that proficient readers in Grade 3 have not yet successfully developed the automatic process. On the other hand, proficient readers in Grades 4-5 display significant advantages in dimensions reflecting both speed and accuracy as compared to non-proficient readers.

To further evaluate the usefulness of categorized correct response times in predicting the criterion and to compare with the continuous response times, stepwise logistic regression models were employed to regress SBAC proficiency classification on scores of different latent traits. The nested model includes only ability as a predictor, whereas the full models include a second predictor; either speed, efficiency, or automaticity. Table 5 shows the results of deviance tests, AUC, AIC, and regression coefficients of each model. The changes of Pseudo (i.e., Nagelkerke) R^2 stand for the increments of Pseudo R^2s of each full model (Models 1-5) over the nested model (Model 0). Across grades, the increments in the Nagelkerke R^2 were consistently the largest for the model that included θ_{au} followed by a model that included θ_e or $\theta_{e.2D}$. Addition of efficiency or automaticity improves the model AIC in each grade. The best fitting model in each grade is that with automaticity as the second predictor. The second best fitting model has either θ_e or $\theta_{e.2D}$ as the second predictor. All of these models are based on dichotomized, *correct* response times rather than continuous response times for all responses. All the models have decent AUC values (above 0.8), suggesting a good validity of discriminating proficient readers and non-proficient readers using the latent trait estimates from the new measure of reading comprehension. Adding a second predictor improved the AUC only in 4th and 5th grades. The deviance test equals the change of -2 Log-Likelihood (deviance) between the nested model (with ability as the only predictor) and a full model (with speed, efficiency, or automaticity as the second predictor). It shows that adding the efficiency score as an additional predictor significantly decreases the deviance, although the effect size in Grade 3 is not as large as those in Grades 4-5. The latent variables τ' performs equally as its counterpart τ; the same goes for $\theta_{e.2D}$ and θ_e. On the other hand, as indicated by the deviance tests and the changes of Pseudo R^2, addition of speed estimates shows no significant improvement in Grade 3. In Grades 4-5,

addition of speed estimates shows some improvements of deviance and Pseudo R^2, but not as much improvement as efficiency or automaticity does in predicting the proficiency criterion. The results in Table 5 tend to support the validity of dimensions based on response times as having incremental validity in their own right. Somewhat surprisingly, they provide the strongest evidence for a dimension based on response times and accuracy, rather than accuracy alone. Dimensions based on dichotomized, *correct* response times receive at least as much support as the dimensions based on continuous response times for all responses.

ROC curves (Figure 5) of the logistic regression models are created to examine the contribution of correct response time information above and beyond the response accuracy. The ROC curve plots the sensitivity (i.e., true positive rate) against 1-specificity (i.e., false positive rate). In Grade 3 (Figures 5a-5d) the ROC curve of the nested model with only ability as the predictor almost overlaps with the ROC curve of a full model with a second predictor of either efficiency, speed, or automaticity. The gap between two ROC curves becomes larger in Grade 4 (Figures 5e-5h) and Grade 5 (Figures 5i-5l). This is consistent with the AUC observed in Table 5, that the AUCs of the full and nested models in Grade 3 are the same whereas in Grades 4-5 the AUC of the nested model is noticeably smaller than the AUC of each full model.

Conclusion and Discussion

The biggest incentive for studying response times is that the gain in the precision and validity might be the same as lengthening the test, but with no additional cost. For example, based on the Spearman-Brown equation, MOCCA forms need additional 15 to 50 items to achieve the same reliabilities as incorporating response times. Response times have been utilized in previous studies as ancillary information in estimating the latent trait of ability. Taking response times into account could result in appreciable changes in the values of the latent ability

estimates with respect to the estimates obtained on the sole basis of responses. Using a carefully designed new measure, this study observes that the latent trait measured by either accuracy or response times are only slightly altered by inclusion of another dimension, if conceived as an ancillary dimension, in a multidimensional model. The response time dimension, conceptualized as a target efficiency dimension in its own right, could improve the prediction of future performance over and above an accuracy dimension alone; even more so for a single dimension of automaticity that combines both response times and accuracy.

Speed and efficiency, and ability and automaticity start to diverge in the upper interval of the latent trait continuum, as shown in Figures 2 and 4. Interpretation of these new variables lead the authors interested in studying the cognitive process of endorsing a correct response, since efficiency and automaticity reflect correct response times. The speed of choosing a correct response denotes a dual process. Based on the dual process theory, the current study posits that the latent trait manifested in both response times and accuracy should be distinct from the latent trait manifested in only accuracy, depending on the cognitive processing required by the test. Increments in response times are usually associated with higher ability in tests requiring controlled processing, while the same increments are associated with lower ability in tests requiring automatic processing. The test of reading comprehension requires automatic processing, therefore a skilled, automatic comprehender should be high in response accuracy and low in response times (i.e., faster) according to the dual process theory.

Following this argument, two questions need to be addressed. The first question concerns the relation between ability, speed, and speed of endorsing a *correct* response – in other words, whether the individuals with higher ability endorse a correct response more quickly than average ability individuals, in the reading comprehension assessment. These latent traits share similarity

and dissimilarity with each other. Graphical representation in Figures 1- 4 helps us understand the relation between latent traits. On average it requires individuals with high ability (estimated with only response accuracy) significantly less time to arrive at a correct response than for individuals with moderate ability. Thus, a skilled comprehender is faster in comprehending the passage correctly than a less-skilled comprehender. This is consistent with Petscher et al.'s (2015) finding that the high ability individuals maintained a stronger representation of fast speed. On the contrary, the difference of *incorrect* response times between high ability individuals and moderate ability individuals are not significant. Also, the speed of endorsing correct responses contains less variability than the speed of endorsing incorrect responses. It implies that response times of correct responses could provide useful information differentiating high ability individuals from the moderate ability individuals. Meanwhile, differentiating high and moderate ability may not improve from inclusion of incorrect response times.

The second question this study is particularly interested in is the usefulness of time data of correct responses – in other words, whether including categorized times of only correct responses would benefit the ability estimates and the measure as much as including continuous times of all responses. The present study uses regression models predicting a future performance to examine which operationalization of ability and speed is associated with higher predictive validity. The increments in precision and validity observed by previous studies were usually not substantial. Molenaar et al. (2015) used the chess ability data, including continuous response times for all responses, and observed the change of R^2 in predicting an external criterion using the hierarchical model was only 0.01 – 0.02 compared against the model without response times. Results of the present study show that by including response times as ancillary information, the improvement of the precision of measurement of ability is minimal. However, the improvement

of validity is substantial, especially when response time is combined with accuracy. All the increments in pseudo-R^2 over and above the accuracy dimension are positive, ranging up to 0.06, suggesting the usefulness of including response time dimensions in enhancing the predictive validity. If using another test that has a strict time limit as the validity indicator, the expected effect of adding response time information could be more substantial. Furthermore, the observed increments associated with estimates using categorized response times of correct responses are always larger than those associated with estimates using continuous response times of all responses. This finding is particularly of interest since it only requires the time data of correct responses. The current study suggests there is little advantage to using continuous times. There is much disagreement as to how to model continuous response times in the literature. Dichotomizing relies on a property of the response times (above or below the median) which does not depend on its distribution, and therefore does not require the distribution be known. Dichotomizing correct responses also prevents us from overly rewarding rapid guessers since fast correct responses receive the same credit regardless of how far they are from cut-offs.

As an example of integrating psychological theory and measurement model, results of this study also enhance our understanding of developmental stages of the comprehension process. According to dual processing theory, the automatic process of reading comprehension starts to emerge after the controlled process, but it remains a question when the emerging starts. The present study shows that incorporation of response times, either as an ancillary, a target or a combined dimension, presents noticeable improvement in predicting future performance of reading proficiency, especially in the higher grades (Grades 4 - 5). This study suggests that proficient readers in higher grades of elementary school start to acquire the automatic process.

This study differed from previous literature in three ways: 1) it extends the literature by examining predictive validity of accuracy and speed in reading comprehension assessment through validation with future performance on an external criterion; 2) it proposes two alternative approaches of using response times, that were often used as ancillary information, and both approaches require less data and provide more useful information in the reading assessment; 3) this study also demonstrates how cognitive theory and measurement models can be integrated to model the sequential process of automatic responses. This study suggests that response times of correct responses can be used as dichotomous variables to imply types of the comprehension process. There are other constructs/domains to which dual processing theory may apply, such as intelligence and problem solving, to which the implications of the present study could be easily generalized. For example, a controlled process of mathematical problem solving is associated with higher ability, therefore a fast response may not indicate a more effective ability.

Limitations and Future Study. There are two limitations of this study. The first one is related to the administration of the MOCCA test, where the instruction per teacher might not be consistent. Students were simply told to choose correct answers. There were no instructions regarding speed, because the scores reported to teachers and students were a function of accuracy only. If response speed is used in scoring performance, examinees should be informed that speed is considered in determining their ability, and the scoring rule should be made explicit to examinees so that they could judge how the speed-accuracy tradeoff might impact their scores. The second limitation is that when real reading data is used, the "truth" is not known, which limits the certainty of the conclusions. In future research, a simulation study could be conducted where the truth is known a priori and then the comparisons between the proposed models are

reported, but the usefulness of such simulation results will depend on how precisely the simulation model mirrors the cognitive processes underlying real data.

Reference

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*(4), 443-459.

Carlson, S., Seipel, B., & McMaster, K. (2014). Development of a new reading comprehension assessment: Identifying comprehension differences among readers. *Learning and Individual Differences, 32*, 40 – 53.

Davison, M. L., Semmes, R., Huang, L., & Close, C. N. (2012). On the reliability and validity of a numerical reasoning speed dimension derived from response times collected in computerized testing. *Educational and Psychological Measurement, 72*(2), 245-263.

De Boeck, P., & Partchev, I. (2012). IRTrees: Tree-based item response models of the GLMM family. *Journal of Statistical Software, 48*(1), 1-28.

Dennis, I., & Evans, J. (1996). The speed-error trade-off problem in psychometric testing. *British Journal of Psychology, 87*, 105–129.

Egloff, B., & Schmukle, S. C. (2002). Predictive validity of an implicit association test for assessing anxiety. *Journal of Personality and Social Psychology, 83*(6), 1441.

Ferrando, P. J., & Lorenzo-Seva, U. (2007). An item response theory model for incorporating response time data in binary personality items. *Applied Psychological Measurement, 31*(6), 525-543.

Fox, J.-P., Klein Entink, R. H., & van der Linden, W. J. (2007). Modeling of responses and response times with the package cirt. *Journal of Statistical Software, 20*, 1–14.

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill:

Insights from a computer-based large-scale assessment. *Journal of Educational Psychology, 106*(3), 608–626.

Loeys, T., Rosseel, Y., & Baten, K. (2011). A joint modeling approach for reaction time and accuracy in psycholinguistic experiments. *Psychometrika, 76*(3), 487-503.

Maris, G., & van der Maas, H. (2012). Speed-accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika, 77*(4), 615-633.

Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. (2015). A bivariate generalized linear item response theory modeling framework to the analysis of responses and response times. *Multivariate behavioral research, 50*(1), 56-74.

Muthén, L. K., & Muthén, B. O. (1998-2015). *Mplus user's guide. 7th edition*. Los Angeles, CA: Muthén & Muthén.

Partchev, I., & De Boeck, P. (2012). Can fast and slow intelligence be differentiated?. *Intelligence, 40*(1), 23-32.

Partchev, I., De Boeck, P., & Steyer, R. (2013). How much power and speed is measured in this test? *Assessment, 20*(2), 242-252.

Petscher, Y., Mitchell, A. M., & Foorman, B. R. (2015). Improving the reliability of student scores from speeded assessments: an illustration of conditional item response theory using a computer-administered measure of vocabulary. *Reading and Writing, 28*, 31 – 56.

Samejima, F. (1969). Estimation of latent trait ability using a response pattern of graded scores. *Psychometrika Monograph*, No. 17.

Semmes, R., Davison, M. L., & Close, C. (2011). Modeling individual differences in numerical reasoning speed as a random effect of response time limits. *Applied Psychological Measurement, 35*(6), 433-446.

Smarter Balanced Assessment Consortium: 2014-2015 Technical Report (2016). Retrieved from http://portal.smarterbalanced.org/library/en2014-15, July 7, 2017.

Su, S. (2017). *Incorporating Response Times in Item Response Theory Models of Reading Comprehension Fluency* (Doctoral dissertation). Retrieved from the University of Minnesota Digital Conservancy, http://hdl.handle.net/11299/190489.

Thissen, D. (1983). Timed testing: An approach using item response theory. In D.J.Weiss(Eds), *New horizons in testing* (pp.179–203). New York: Academic Press.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics, 31*(2), 181-204.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72*, 287-308.

van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement, 46*, 247-272.

van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika, 73*, 365–384.

van der Linden, W.J., Scrams, D.J., & Schnipke, D.L. (1999). Using Response-Time Constraints to Control for Differential Speededness in Computerized Adaptive Testing. *Applied Psychological Measurement, 23*(3), 195-210.

Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology, 68*(3), 456-477.

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement, 43*(1), 19-38.

Table 1

Sample Size per Form and Aggregated by Grade

	Grade 3			Grade 4			Grade 5		
Form	3.1	3.2	3.3	4.1	4.2	4.3	5.1	5.2	5.3
	538	518	520	519	500	478	436	395	384
MOCCA		1576			1497			1215	
SBAC		245			426			386	

Table 2

Correlation between Latent Trait Estimates in Different Models

		θ_a	τ	θ_e	θ_{au}	θ'_a	τ'	$\theta_{a.2D}$	$\theta_{e.2D}$
Grade 3	θ_a	1							
(N=1576)	τ	0.30	1						
	θ_e	-0.06	-0.77	1					
	θ_{au}	0.86	0.02	0.34	1				
	θ'_a	1	0.33	-0.09	0.84	1			
	τ'	0	0.95	-0.79	-0.24	0.04	1		
	$\theta_{a.2D}$	1	0.31	-0.07	0.85	1	0.01	1	
	$\theta_{e.2D}$	-0.07	-0.77	1	0.33	-0.10	-0.79	-0.08	1
Grade 4	θ_a	1							
(N=1497)	τ	0.30	1						
	θ_e	0.01	-0.72	1					
	θ_{au}	0.84	0	0.44	1				
	θ'_a	1	0.33	-0.02	0.83	1			
	τ'	0.01	0.96	-0.76	-0.25	0.04	1		
	$\theta_{a.2D}$	1	0.30	0.01	0.84	1	0.01	1	
	$\theta_{e.2D}$	0.02	-0.72	1	0.44	-0.02	-0.75	0.02	1
Grade 5	θ_a	1							
(N=1215)	τ	0.25	1						
	θ_e	0.1	-0.69	1					
	θ_{au}	0.82	-0.06	0.55	1				
	θ'_a	1	0.27	0.07	0.80	1			
	τ'	0.01	0.97	-0.73	-0.26	0.03	1		
	$\theta_{a.2D}$	1	0.23	0.11	0.83	1	0	1	
	$\theta_{e.2D}$	0.11	-0.69	1	0.56	0.09	-0.73	0.13	1

Note. θ_a is the ability dimension of the unidimensional model; τ is the speed dimension of the lognormal model; θ_e is the efficiency dimension of the unidimensional model; θ_{au} is the automaticity estimate of the GRM; θ'_a is the ability dimension of the hierarchical model; τ' is the speed dimension of the hierarchical model; $\theta_{a.2D}$ is the ability dimension of the 2DTREE model; $\theta_{e.2D}$ is the efficiency dimension of the 2DTREE model.

Table 3

Average Scaled Conditional Standard Error for Each Form

	Grade 3			Grade 4			Grade 5		
Form	3.1	3.2	3.3	4.1	4.2	4.3	5.1	5.2	5.3
θ_a	0.31	0.32	0.32	0.33	0.33	0.33	0.34	0.35	0.33
θ'_a	0.31	0.32	0.32	0.33	0.33	0.33	0.34	0.35	0.33
$\theta_{a.2D}$	0.31	0.32	0.33	0.33	0.33	0.33	0.34	0.35	0.33

Table 4

Average Latent Trait Estimates of Proficient and Non-Proficient Students

	Grade 3				Grade 4				Grade 5			
	Prof.	Non-Prof.	*T*	*d*	Prof.	Non-Prof.	*T*	*d*	Prof.	Non-Prof.	*T*	*d*
θ_a	0.74	-0.45	-13.29**	1.70	0.80	-0.36	-16.28**	1.59	0.77	-0.21	-12.62**	1.30
θ_e	-0.09	-0.25	-1.40	0.18	0.27	-0.38	-6.97**	0.67	0.34	-0.28	-6.59**	0.67
θ_{au}	0.63	-0.50	-12.13**	1.56	0.82	-0.45	-16.20**	1.56	0.77	-0.28	-12.87**	1.31
τ	0.34	0.13	-1.88	0.24	0.06	0.31	2.96**	0.29	0.03	0.18	1.71	0.18
θ'_a	0.75	-0.44	-13.17**	1.68	0.79	-0.33	-15.91**	1.55	0.77	-0.20	-12.34**	1.28
τ'	0.12	0.26	1.37	0.17	-0.17	0.41	7.32**	0.72	-0.16	0.23	4.70**	0.49
$\theta_{a.2D}$	0.74	-0.45	-13.25**	1.69	0.80	-0.35	-16.32**	1.59	0.78	-0.21	-12.77**	1.32
$\theta_{e.2D}$	-0.09	-0.24	-1.29	0.17	0.27	-0.37	-6.99**	0.68	0.35	-0.28	-6.72**	0.69

Note. Prof. = Proficient in SBAC reading; Non-Prof. = Non-Proficient in SBAC reading.

Table 5

Stepwise Logistic Regression Models of SBAC Proficiency Classification of Grades 3-5

	Model	Coefficients	AUC	AIC	Deviance Test	Change of Pseudo R^2
Grade 3	Model 0: Y ~ θ_a	b_0=-0.32, b_1=2.08**	0.88	220.65	122.89**	0.526
(N=245)	Model 1: Y~θ_a+θ_e	b_0=-0.20, b_1=2.15**, b_2=0.41*	0.88	218.52	4.13*	0.013
	Model 2: Y~θ_a+$\theta_{e.2D}$	b_0=-0.20, b_1=2.16**, b_2=0.40*	0.88	218.58	4.07*	0.013
	Model 3: Y~θ_a+ τ	b_0=-0.24, b_1=2.17**, b_2=-0.24	0.88	221.38	1.27	0.004
	Model 4: Y~θ_a+τ'	b_0=-0.25, b_1=2.10**, b_2=-0.24	0.88	221.46	1.19	0.004
	Model 5: Y~θ_a+θ_{au}	b_0=-0.23, b_1=1.15**, b_2=1.36**	0.88	212.39	10.26**	0.033
Grade 4	Model 0: Y ~ θ_a	b_0=-0.43**, b_1=1.94**	0.86	399.65	194.45**	0.489
(N=426)	Model 1: Y~θ_a+θ_e	b_0=-0.26, b_1=1.95**, b_2=0.73**	0.88	373.34	28.31**	0.054
	Model 2: Y~θ_a+$\theta_{e.2D}$	b_0=-0.26, b_1=1.95**, b_2=0.73**	0.88	373.37	28.28**	0.054
	Model 3: Y~θ_a+ τ	b_0=-0.29*, b_1=2.16**, b_2=-0.78**	0.88	376.50	25.16**	0.048
	Model 4: Y~θ_a+τ'	b_0=-0.29*, b_1=1.93**, b_2=-0.80**	0.88	376.61	25.05**	0.048
	Model 5: Y~θ_a+θ_{au}	b_0=-0.36**, b_1=0.95**, b_2=1.41**	0.88	371.06	30.59**	0.059
Grade 5	Model 0: Y ~ θ_a	b_0=-0.39**, b_1=1.60**	0.81	407.24	130.61**	0.383
(N=386)	Model 1: Y~θ_a+θ_e	b_0=-0.34*, b_1=1.54**, b_2=0.66**	0.84	385.01	24.23**	0.058
	Model 2: Y~θ_a+$\theta_{e.2D}$	b_0=-0.34*, b_1=1.53**, b_2=0.66**	0.84	385.09	24.15**	0.058
	Model 3: Y~θ_a+ τ	b_0=-0.34*, b_1=1.73**, b_2=-0.58**	0.83	395.33	13.91**	0.034
	Model 4: Y~θ_a+τ'	b_0=-0.34*, b_1=1.59**, b_2=-0.59**	0.83	395.35	13.90**	0.034
	Model 5: Y~θ_a+θ_{au}	b_0=-0.43**, b_1=0.87**, b_2=1.08**	0.84	384.74	24.50**	0.059

Note. Y is the criterion, SBAC proficiency classification, in each model. b_0 is the coefficient of intercept; b_1 is the coefficient of ability estimate as the 1st predictor; b_2 is the coefficient of efficiency/automaticity/speed estimates as the 2nd predictor. Deviance test reflects the change of -2LL of the full model against the nested model. The deviance of Model 0 and Model 0' is computed by comparing against the null model with only intercept. Change of pseudo R^2 compares Nagelkerke's pseudo R^2 of the full model against the nested model. $p < .05$, *; $p < .01$, **.

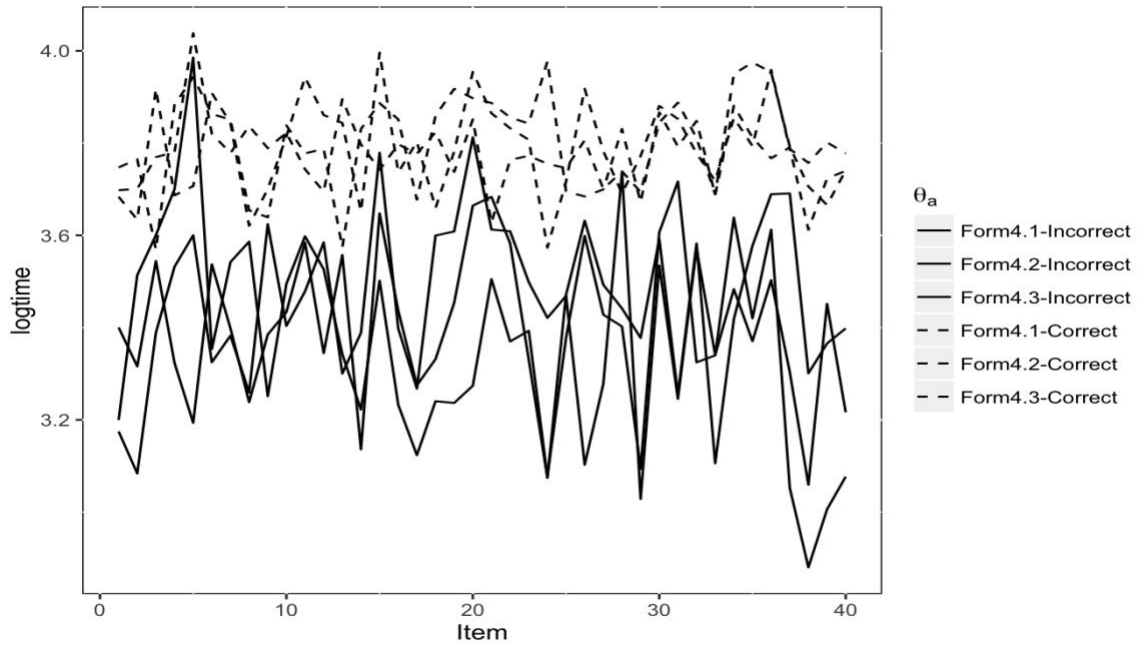

Figure 1. Log response time of incorrect and correct respondents across items in Grade 4

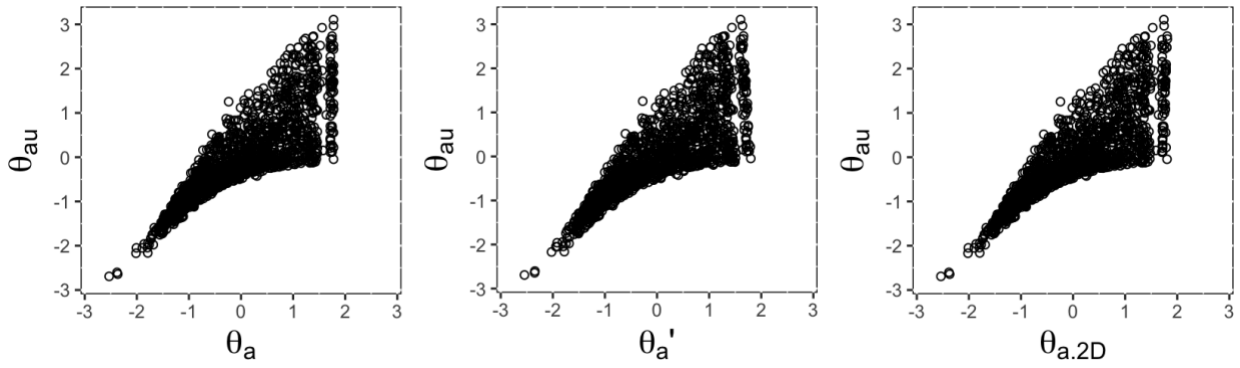

Figure 2. Scatter Plots of Latent Estimates of ability dimensions vs automaticity of Grade 4

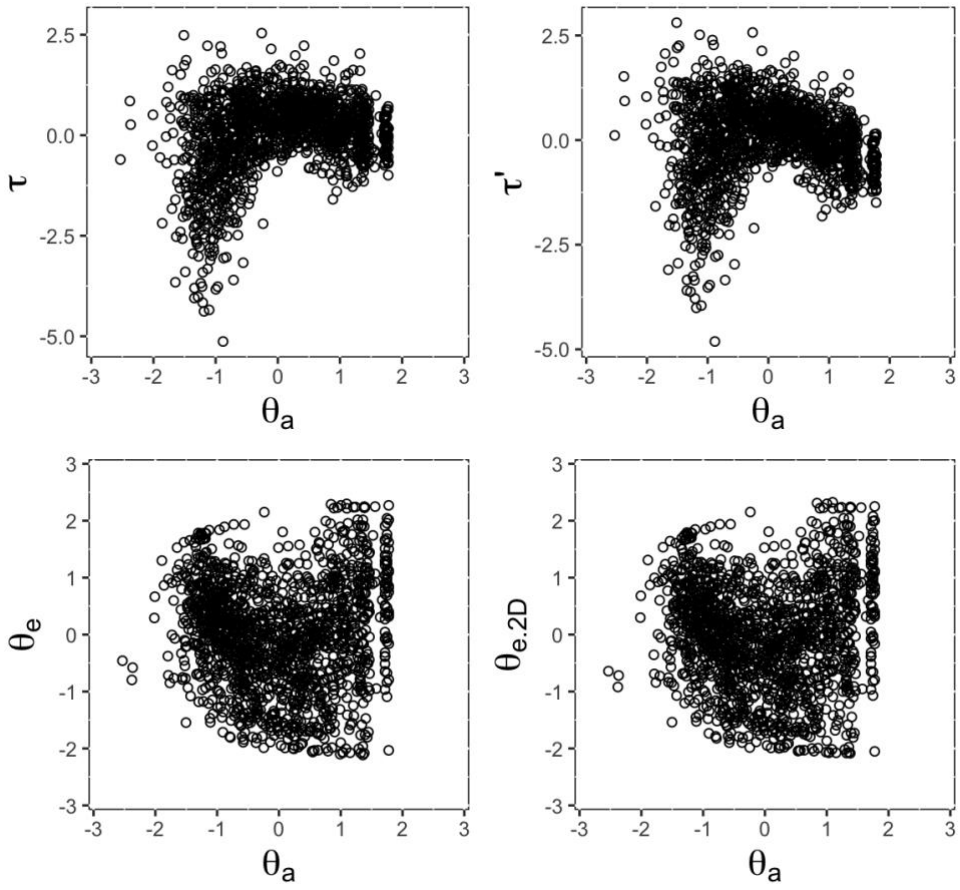

Figure 3. Scatter plots of latent estimates of ability vs speed and ability vs efficiency of Grade 4

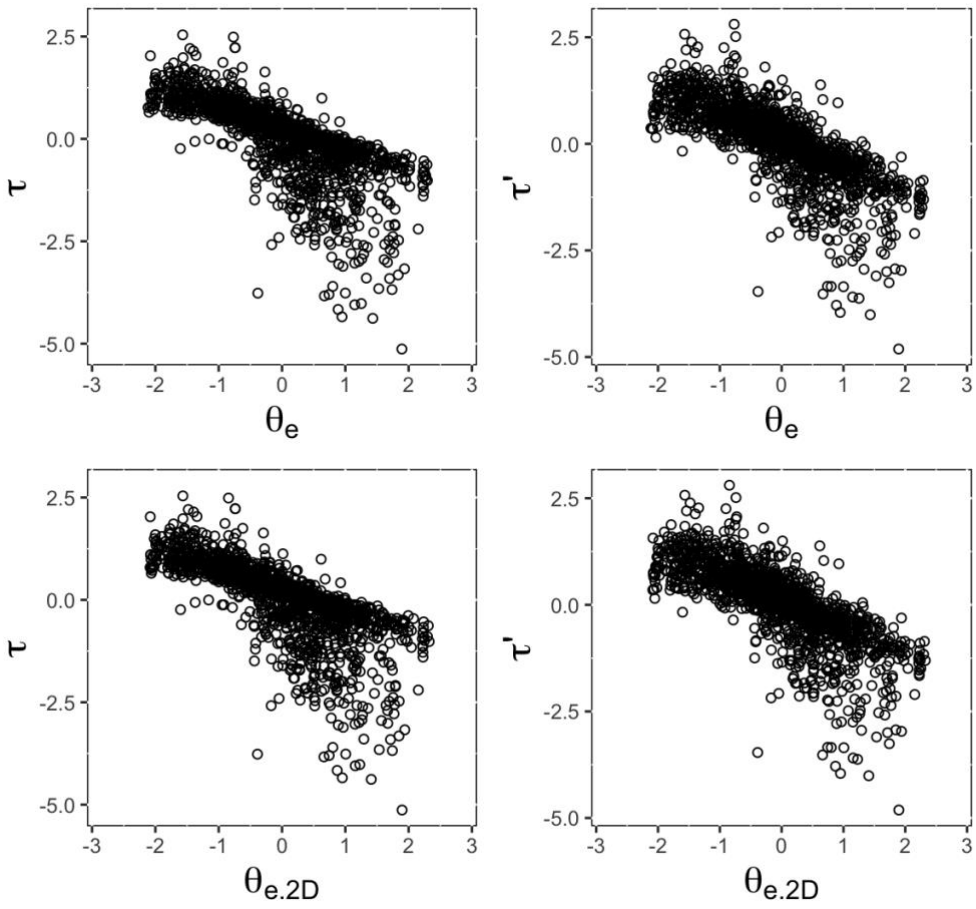

Figure 4. Scatter plots of latent estimates of efficiency vs speed of Grade 4

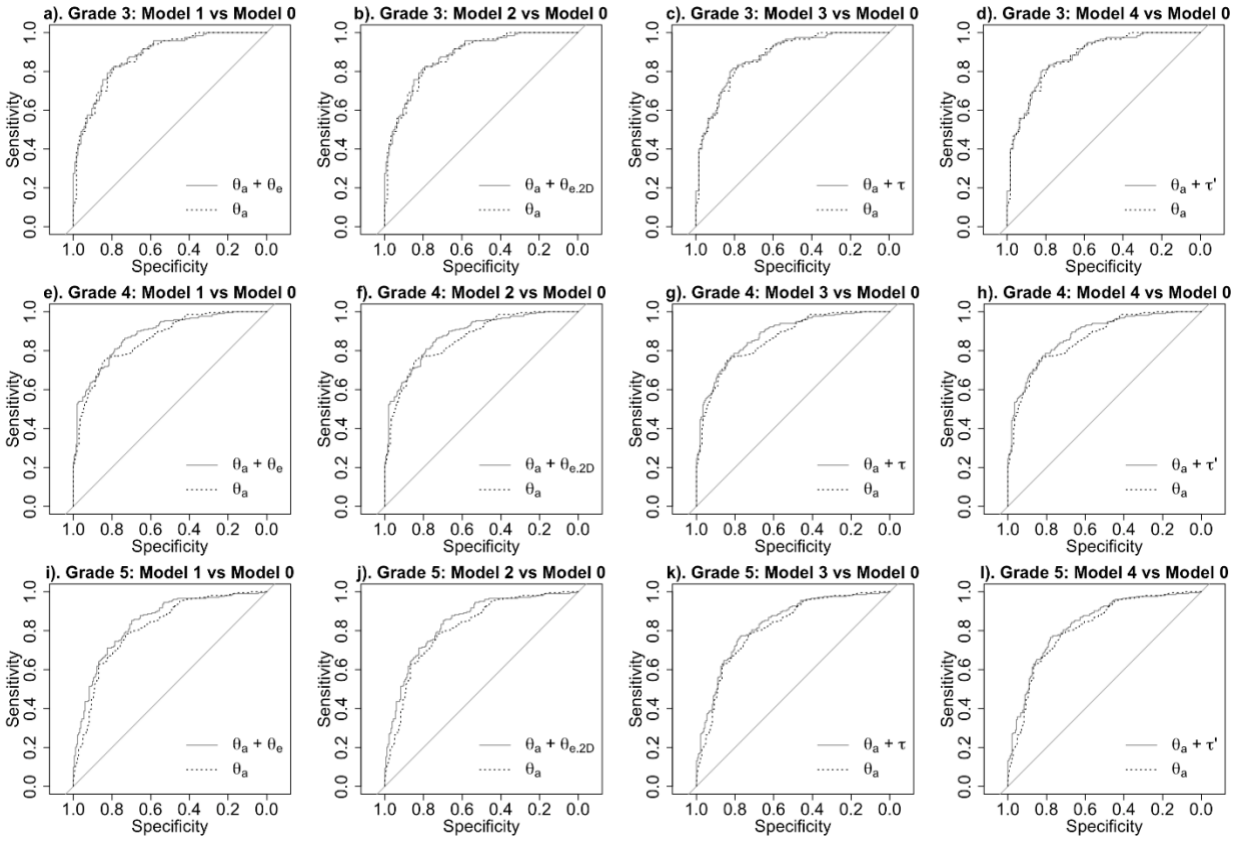

Figure 5. ROC curves of Logistic Regression Models for Grade 3 – 5